计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):267−272 [doi: 10.15888/j.cnki.csa.008307]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

BERT 与 GSDMM 融合的聚类短文本分类①

刘　豪[1], 王雨辰[2]

[1](中国科学技术大学 管理学院 统计与金融系, 合肥 230041)
[2](中国科学技术大学 管理学院 国际金融研究院, 合肥 230041)
通信作者: 刘　豪, E-mail: lh1@mail.ustc.edu.cn

摘　要: 在文本分类任务中, 由于短文本具有特征稀疏, 用词不规范等特点, 传统的自然语言处理方法在短文本分类中具有局限性. 针对短文本的特点, 本文提出一种基于 BERT (bidirectional encoder representations from Transformers) 与 GSDMM (collapsed Gibbs sampling algorithm for the Dirichlet multinomial mixture model) 融合和聚类指导的短文本分类算法, 用以提高短文本分类有效性与准确性. 本算法一方面通过 BERT 与 GSDMM 融合模型将短文本转化为集成语义向量, 集成的向量体现了全局语义特征与主题特征, 解决了短文本特征稀疏与主题信息匮乏的问题. 另一方面在分类器前端训练中通过引入聚类指导算法实现对标注数据的扩展, 同时也提升了结果的可解释性. 最后利用扩展后的标注数据集训练分类器完成对短文本的自动化分类. 将电商平台的差评数据作为验证数据集, 在多组对比实验中验证了本算法在短文本分类方面应用的有效性与优势.

关键词: GSDMM; BERT; SVM; 短文本分类; 聚类指导; 语义向量

引用格式: 刘豪,王雨辰.BERT 与 GSDMM 融合的聚类短文本分类.计算机系统应用,2022,31(2):267–272. http://www.c-s-a.org.cn/1003-3254/8307.html

Clustering Short Text Classification Based on Fusion of BERT and GSDMM

LIU Hao[1], WANG Yu-Chen[2]

[1](Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230041, China)
[2](International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230041, China)

Abstract: In the task of text classification, traditional natural language processing methods have limitations in short text classification due to the sparse features and irregular wording of short texts. Considering the characteristics of short texts, this study proposes a classification algorithm based on the fusion of bidirectional encoder representations from Transformers (BERT) and a collapsed Gibbs sampling algorithm for the Dirichlet multinomial mixture model (GSDMM) and clustering guidance to improve the effectiveness and accuracy of short text classification. First, the model converts short texts into integrated semantic vectors by using the fusion model of BERT and GSDMM. The integrated vectors reflect global semantic features and topic features and solve the problems of sparse short text features and the lack of topic information. Then, the clustering guidance algorithm is introduced into the front-end training of the classifier, which realizes the expansion of the labeled data and improves the interpretability of the results. Finally, the expanded labeled data set is used to train the classifier to complete the automatic classification of short texts. Taking the negative comment of an e-commerce platform as the verification data set, this study verifies the effectiveness and advantages of the algorithm in short text classification in multiple groups of comparative experiments.

Key words: collapsed Gibbs sampling algorithm for the Dirichlet multinomial mixture model (GSDMM); bidirectional encoder representations from Transformers (BERT); SVM; short text classification; clustering guidance; semantic vector

① 基金项目: 安徽省自然科学基金青年项目 (1908085AG299)
收稿时间: 2021-04-06; 修改时间: 2021-04-30; 采用时间: 2021-05-28; csa 在线出版时间: 2022-01-17

随着互联网经济的快速发展, 在各个线上线下平台产生了大量包括文本在内的数据, 对于上述数据的处理分析具有重要的意义. 不同于其他类型数据, 文本数据需要深度的理解分析, 依赖简单的统计方法难以有效处理分析此类数据. 如何实现文本数据的自动化处理与分析成为亟待解决的问题. 而随着自然语言处理技术的成熟, 相关技术也为分析文本数据提供了可行有效的思路和工具. 在这些自然语言处理技术中, 文本分类是一个关键的和基本的技术, 在各种下游应用中具有重要的作用. 对于文本的分类分析问题, 此前相关的研究主要都针对相对正式的规范性长文本数据[1–3]. 虽然 Lv 等[4] 也对短文本数据分类提出相应的处理方法, 但在特征提取方面没有考虑短文本的特性, 实际表现不佳.

不同于规范性的长文本数据, 短文本数据具有以下典型特点.

1) 稀疏性, 每条短文本数据的字符长度都比较短, 一般都在 200 字以内, 因此文本所包含的有效信息少, 造成特征稀疏, 并且特征集的维数非常高, 很难抽取到关键样本特征用于分类学习.

2) 数据量大, 更新快.

3) 用词不规范, 形式不统一, 噪声特征多, 依靠单纯的统计分析很难得到实际的语义信息.

传统文本处理算法从早期的浅层统计学习模型, 例如朴素贝叶斯, K 近邻等算法对于上下文信息的理解与全局语义信息的利用十分有限. 而近几年发展迅速的深度学习算法则很好地克服了上述缺点, 例如 Hochreiter 等[5] 于 1997 年提出长短记忆神经网络 (LSTM) 针对 RNN 算法的梯度爆炸与梯度消失现象, 通过在神经元中设置门结构, 可以对文本中的重要信息选择性地进行长期记忆, 更好地融合上下文信息. 但是, 相较于传统长文本数据, 微博、商品差评、推特等短文本数据具有特征维度稀疏、缺乏上下文信息、用词不规范等特点, 这些特点使得深度学习模型在短文本任务上表现不佳. 而 Devlin 等[6] 在 2018 年提出的 BERT 模型则进一步提升了准确率与效率, 通过利用前后两方向信息, 基于 Transformer 模型引入 Mask 训练方式加强了对于上下文的语义理解, 还在应用了大型语料库预训练机制, 加入了对预训练模型的微调技术. BERT 通过上述技术在自然语言处理多个领域应用取得成功. 正因为 BERT 模型的在文本处理方面的良好效果, 所以本文选择基于 BERT 模型对短文本进行全局语义特征提取. 此外, Peinelt 等[7] 在计算句子相似度任务中提出的主题特征结合全局语义特征思想, 特别的, Peinelt 等提到 BERT 模型添加主题信息有助提高针对特定行业板块与知识领域的表现, 所以本文认为主题信息融合 BERT 模型得到的全局语义信息可能会有助于提高算法对短文本分类任务的处理性能. 国内学者也有将主题与全局语义特征拼接的尝试, 付静等[8] 在 2021 年将隐层狄利克雷 (latent Dirichlet allocation, LDA) 模型与 BERT 模型融合形成拼接向量, 相较于传统词向量方法与单独的 BERT 模型, 显示出了更优良的语义表达性能. 但是他们未能考虑到 LDA 对于短文本处理的适用性[9]. 本文针对短文本特点选择 Yin 等[10] 在 2014 年提出的 GSDMM 算法作为主题分类的模型, GSDMM 算法是一种基于狄利克雷多项式混合模型 (DMM) 的折叠型吉布斯采样算法[11]. GSDMM 与之前的方法不同之处一方面在于算法本身不需要文本的空间向量表达, 而是直接对文档和词进行概率估计, 故可以有效解决文本数据的高维和稀疏问题. 另一方面, GSDMM 文本单个主题的假设也更加符合短文本的特征. 在训练阶段本文则引入集成语义向量聚类指导[12], 利用聚类指导在标注训练集上进行扩展, 提高分类器的训练有效性, 同时也提高了对分类结果的解释性.

通过引入以上模型, 本文针对短文本分类, 提出了 BERT 与 GSDMM 融合聚类的分类算法, 如算法 1.

算法 1. BERT 与 GSDMM 融合聚类的分类算法

1) 基于 BERT 得到短文本总体语义特征向量.
2) 基于 GSDMM 得到短文本的主题向量.
3) 拼接总体语义向量与主题向量形成集成向量.
4) 通过聚类指导扩展训练集.
5) 扩展训练集结合前期标注训练 SVM 分类器.

本文在考虑评价短文本特性的基础上通过自然语言处理技术实现了对短文本数据有效的自动化分类, 整体的流程图如图 1 所示.

1 语义向量嵌入拼接

语料收集完成之后, 需要对文本进行预处理. 文本预处理主要是切分词、去停用词等. 本文使用 Jieba 作为分词工具, 使用公开的中文停用词字典作为标准. 特征提取分为两部分, 两部分分别得到总体语义特征和主题特征. 在进行分类器训练前将两部分的特征向量

进行拼接. 在这一步骤还需要对训练集进行标注, 形成评价分类的训练集W_{train}, 为后续步骤提供监督学习的标签样本. 具体的标注细节在后面章节介绍.

1.1 BERT 算法

BERT 即基于 Transformers 模型得来的双向编码表征模型. BERT 主要有预训练 (pretraining) 与微调 (fine-tuning) 两个步骤. BERT 的关键技术创新是基于 Transformer (Vaswani 等[13] 在 2017 年提出的一种注意力模型) 的双向训练应用于语言建模. 这与以前的方法有所不同, 之前的方法按照从左到右或从左到右和从右到左的顺序训练查看文本序列. BERT 的结果表明, 经过双向训练的语言模型比单向语言模型具有更深的全局语义特征. 在 BERT中双向训练模型正是通过 Masked LM (MLM) 这种新技术在以前不可能的模型中进行双向训练. 具体来说, Mask 技术在将单词序列输入 BERT 之前, 每个序列中 15% 的单词被替换为 Mask. 然后该模型将根据序列中其他未屏蔽单词提供的上下文, 尝试预测被屏蔽单词的原值. BERT 中的另一重要技术是可以将句子对作为模型训练的输入, 并学习预测成对的第 2 个句子是否是原始文档中的后续句子. 在训练期间, 输入的 50% 是一对, 其中第 2 个句子是原始文档中的后续句子, 而在其他 50% 的输入中, 从语料库中选择一个随机句子作为第 2 个句子. 假定随机句子将与第 1 句断开.

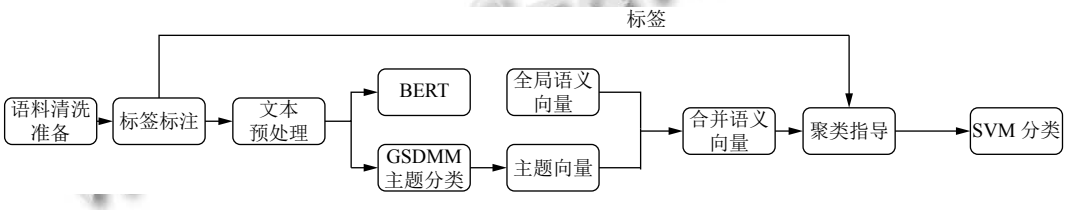

图 1 短文本分类处理流程图

基于 BERT 模型的预训练特点, BERT 模型可以根据下游任务的需要, 选择相应的微调方法进行应用, 这些下游任务包括了问题回答、句子分类和句子对回归. 但是 BERT 模型本身并不支持独立句子嵌入运算. 之前的研究一般使用平均单词嵌入或者使用特殊 CLStoken 的方法试图解决这一问题. 但是结果并不理想[14–16].

本文中采用了 Reimers 等[17] 提出的 Sentence-BERT 方法在预训练模型上微调实现独立句子嵌入. 选择 Sentence-BERT 方法的主要原因有如下两点:

1) Sentence-BERT 通过引入 Siamese 和 Triplet 网络[18] 更新权重使得独立句子嵌入具有语义上的意义, 而具有语义信息的向量可以用于后续的聚类任务.

2) Sentence-BERT 在 BERT 的输出中加入了一个汇总的操作 (平均池化), 可以导出一个固定 768 维的句子嵌入向量, 便于的后续分析运算.

通过 Sentence-BERT 方法可以完成对文本全局语义信息的特征提取, 提取的特征向量可以用于下一阶段集成拼接.

1.2 GSDMM 算法

GSDMM 是一种无监督文本主题聚类模型, 该模型根据狄利克雷多项式混合模型 (Dirichlet multinomial mixture model, DMM) 生成文档, 利用折叠吉布斯采样 (Gibbs sampling) 近似求解模型. 模型假定文档是根据混合多项式分布产生的, 并且主题和文档之间相互对应. GSDMM 基于完备性与一致性两大原则实现聚类, 完备性即聚类处理的所有文本都会被分到具体的簇中, 而一致性则指的是被聚类到同一个簇的文本都尽可能的相似.

GSDMM 算法以 Nigam 等[19] 于 2000 提出的 DMM 概率生成模型为基础. DMM 模型结构与 LDA 类似. DMM 模型结构相对于 LDA 的改进在于假设每个文档只含有一个主题[20], 而 LDA 则假定每个文档含有多个主题. 单个主题的模型假定更加符合短文本实际情况. DMM 模型中变量定义如表 1.

文档 d 产生过程可以表示如下:

首先根据主题 (聚类簇) 的权重$p(z=k)$选择一个主题 k. 然后根据$p(d/z=k)$分布选择主题生成文档 d. 因此, 我们可以用所有主题的概率总和来描述文档 d 的可能性:

$$p(d)=\sum_{k=1}^{k=K} p(d/z=k)p(z=k) \tag{1}$$

其中, K 指的是主题的数量, 在得到文档生成概率公式后, 接下来的问题关键在于求得$p(d/z=k)$与$p(z=k)$ 表达方式.

在朴素贝叶斯假定下$p(d/z=k)$可以表达为:

$$p(d/z=k)=\prod_{w\in d}p(w/z=k) \tag{2}$$

表 1 DMM 模型中变量定义

变量	定义
K	主题个数
D	语料库中文档个数
L	文档平均长度
d	语料库中的文档
z	每个文档的主题
V	单词数
I	迭代次数
m_z	主题z中文档数
n_z	主题z中的字数
n_z^w	单词w在簇z中出现的次数
N_d	文档d中的词数
N_d^w	文档d中单词w出现的次数

模型中假定主题在单词上是多项式分布:

$$p(w\mid z=k)=p(w\mid z=k,\Phi)=\phi_{k,w} \tag{3}$$

其中, $\phi_{k,w}$指主题 k 的词分布, 在文档 d 中, $\sum_{w=1}^{V}\phi_{k,w}=1$ (V 表示文档 d 的词向量维度), Φ是服从概率分布的变量. 狄利克雷分布式是多项式分布的共轭先验概率分布, 选择以β为参数的狄利克雷分布作为Φ先验分布:

$$p(\Phi\mid\vec{\beta})=Dir\left(\vec{\phi}_k\mid\vec{\beta}\right) \tag{4}$$

每个主题的概率服从多项分布:

$$p(z=k)=p(z=k/\Theta)=\theta_k \tag{5}$$

其中, Θ是主题分布矩阵, θ_k表示主题分布, 在文档 d 中, $\sum_{w=1}^{V}\theta_k=1$. 选择以α为参数的狄利克雷分布作为Θ的先验分布. 则可以得到下式:

$$p(\Theta\mid\vec{\alpha})=Dir(\vec{\theta}\mid\vec{\alpha}) \tag{6}$$

得到表达式之后, 需要利用折叠吉布斯抽样算法得到每篇文档所属的主题的集合, 即文档的主题 z. Gibbs 采样的物理过程, 实际上就是一个词在不同的主题上不断地采样, 最终得到这个词的主题分布矩阵, 从而得到文档的主题分布和主题的词分布. 利用 Gibbs 采样法对模型进行求解, 在训练过程中采样的一篇文档属于某个主题的概率如下:

$$p\left(z_d=z\mid\vec{z}_{\neg d},\vec{d}\right)\propto\frac{m_{z,\neg d}+\alpha}{D-1+K\alpha}\frac{\prod_{w\in d}\prod_{j=1}^{N_u^d}\left(n_{z,\neg d}^w+\beta+j-1\right)}{\prod_{i=1}^{N_w^d}\left(n_{z,\neg d}+V\beta+i-1\right)} \tag{7}$$

其中, $\neg d$表示去除当前文档 d 的信息.

GSDMM 模型的求解过程如下:

(1) 初始化模型参数: K, α, β, I; 初始化变量$m_z=n_z=n_z^w=0$.

(2) 获得文档集$\vec{d}$, 对于每个文档 d. 一方面赋予一个服从多项分布的随机主题: $z_d\leftarrow z\sim$ Multi(1/K). 另一方面更新变量: $m_z=m_z+1$, $n_z=n_z+N_d$, 对于该文档的每个词 w 有 $n_z^w=N_d+N_d^w$.

对文档集中的所有文档初始化完成后, 得到 K 个属于不同主题的集合, 且每个文档只属于一个主题.

(3) 进行 Gibbs 采样, 对每个文档 d:

1) 记录当前 d 所属的主题: $z=z_d$.

2) 当前主题中去除 d 的信息:

$m_z=m_z-1$, $n_z=n_z-N_d$, 对该文档包含的词 w 有 $n_z^w=N_d-N_d^w$.

3) 根据条件分布为 d 重新分配主题:

$$z_d\leftarrow z\sim p\left(z_d=z\mid\vec{z}_{\neg d},\vec{d}\right)$$

4) 更新变量:

$$m_z=m_z+1,n_z=n_z+N_d,n_z^w=N_d+N_d^w$$

(4) 重复过程 (3), 直到最大迭代次数 I.

(5) 输出每个文档的类别标签.

通过上述方法, GSDMM 能够处理文本数据的高维和稀疏问题得到文档的类别标签, 为下一步集成语义向量提供主题信息.

1.3 语义向量拼接

在得到句子的全局语义特征与主题特征后, 使用向量拼接的方式完成主题粒度下对全局语义特征的扩充, 如图 2 中所示. cls 符号在输入单独句子时插入, 其对应的输出向量可以作为整篇文本的语义表示, 用于文本分类. Token 1 作为文档中字符的表达, E-token 1 代表了 Embedding (嵌入) 层中 token 1 对应的特征. T1 表示经过 BERT 处理之后的输出特征, 以此类推.

2 聚类指导

由于评价文本的标注需要耗费大量的人力与精力, 所以如何能够利用有限的标注数据训练分类器成为了文本分类的重要问题. 本文通过引入聚类指导的方法, 对上一阶段生成的集成语义向量进行聚类. 训练集中未标注的数据标注为同一簇族内最大数量标签的属性. 同时考虑到聚类算法的有效性, 需要根据轮廓系数对于

聚类的结果进行讨论, 确定最终的标注. 此外, 语义向量上的聚类特征也可以帮助电商平台理解分析不同主题, 聚类特征被表示为从聚类语义向量中揭示的潜在主题. 综合来说, 聚类指导的加入在以下两方面提高算法的性能与实用性: (1) 只需要少量标注数据就可以有效实现分类器. (2) 提供了短文本之间相似性的直观解释, 这将有利于语义理解和标注的讨论. 本文使用经典的基于空间密度的 DBSCAN 算法对语义向量进行聚类.

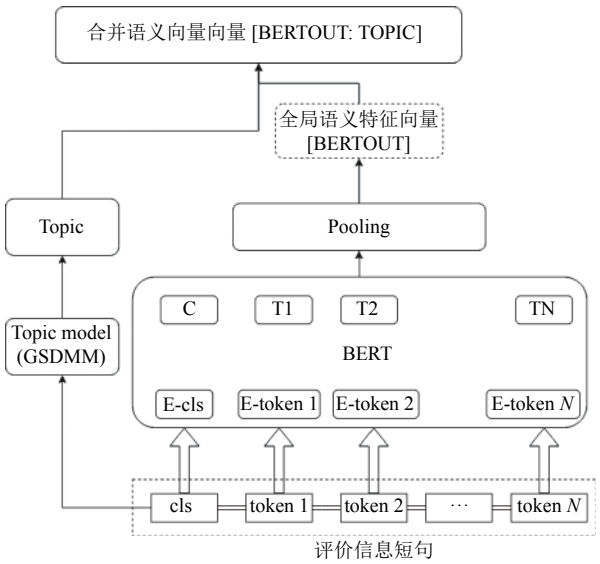

图 2 语义向量拼接示意图

聚类指导形成扩展训练集C_{train}, 为下一步训练短文本分类模型做准备.

3 分类识别

此步骤通过前期标注分类的训练集W_{train}与聚类步骤得到的扩展标注集C_{train}, 利用W_{train}与C_{train}标签训练 SVM 分类器. 当训练完成 SVM 分类器后, 通过每个评价的集成语义向量就可以得到相应的分类标签.

4 实验结果

本文提出的算法为短文本分类提供了可行有效的思路和工具. 为验证文中分类方法的有效性, 本文选取了某电商平台上 150 种不同类型产品, 特定产品下随机抽取 200 条客户差评, 总计收集了 30 000 条文本. 其中 18 000 条文本作为训练集, 12 000 条文本作为测试集. 在 18 000 条训练文本中人工标注分类 6 000 条文本, 差评文本的长度分布如表 2 所示.

表 2 电商平台差评文本长度分布

字数 (个)	数量
1–20	16 290
21–100	13 530
101–194	180

标注的标签类型按照电商平台给出的差评标签通过筛选总结分为相关服务、价格、物流发货、产品质量、无效恶意差评 5 个种类.

表 3 是采用不同模型提取语义向量利用测试集得出的实验结果. 在模型对比中所有参数一致, 分类器统一采用 SVM. BERT+LDA 的方法参照付静等[8] 研究. 多分类问题按照指标平均数取值. 在表 3 中可以看出 GSDMM 与 BERT 融合模型在 3 个指标上均优于其他模型, 这一结果可能源于 GSDMM处理短文本数据的良好性能与语义向量的集成的效果. 基于本算法未来在这方面可以通过扩大训练数据集与参数调优的方法加以进一步提升.

表 3 模型表现对比表 (%)

模型类型	准确率	召回率	*F*1值
BERT+GSDMM	81.2	83.3	82.2
BERT+LDA	79.6	80.1	79.8
BERT	76.3	77.8	77.1

5 结论与展望

本文针对短文本分类问题提供了有效的解决算法. 之前的文献研究中出现过采用 BERT 模型以及采用 GSDM 算法的短文本聚类分类算法. 然而, 本文的价值在于将二者相结合的应用探索. 本算法通过融合 GSDMM 与 BERT 有效提取短文本的主题与总体语义特征, 这一方法在一定程度上解决了短文本数据缺乏上下文信息与主题信息的问题. 同时在分类器训练前端引入聚类指导, 提高了分类器的训练效率, 利用少量的标注数据就可以实现短文本的有效分类. 通过将该模型与 BERT 以及 BERT+LDA 模型进行对比实验, 实验结果验证了本文所提方法的可行性和有效性. 本文所提出的算法虽然达到了较好的性能, 但也存在一定的局限性. 例如, 算法模型的性能会受到超参数、数据集选取等因素的影响, 但本文暂时未对这些因素进行综合探究, 这也是我们未来工作的方向之一. 同时聚类算法与分类器下一步也需要改进, 以提高分类的准确度与效率. 总体来说, 在实验验证中本算法有效的实现了短文本自动分

类的功能, 可以通过对电商平台评论的处理分析为未来产品服务的改进提供有效的信息支持, 具有较为广阔的应用场景.

参考文献

1 杨恒, 颜宏文. 基于 DBM 的电力投诉工单分类的应用研究. 计算技术与自动化, 2020, 39(3): 86–90.

2 胡菊香, 吕学强, 刘克会. 利用类别引导词的投诉文本分类利用类别引导词的投诉文本分类. 现代图书情报技术, 2015, (7): 97–103.

3 北京工业大学. 一种基于 SVM 参数优化的投诉举报文本分类方法: 中国, 111753083A. 2020-10-09.

4 Lv GY, Xu T, Chen EH, *et al*. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: ACM, 2016. 3000–3006.

5 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

6 Devlin J, Chang M W, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186.

7 Peinelt N, Nguyen D, Liakata M. tBERT: Topic models and BERT joining forces for semantic similarity detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 7047–7055.

8 付静, 龚永罡, 廉小亲, 等. 基于 BERT-LDA 的新闻短文本分类方法. 信息技术与信息化, 2021, (2): 127–129. [doi: 10.3969/j.issn.1672-9528.2021.02.044]

9 Hong LJ, Davison BD. Empirical study of topic modeling in twitter. Proceedings of the First Workshop on Social Media Analytics. Washington: ACM, 2010. 80–88.

10 Yin JH, Wang JY. A dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014. 233–242.

11 Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, PAMI-6(6): 721–741. [doi: 10.1109/TPAMI.1984.4767596]

12 Raskutti B, Ferrá H, Kowalczyk A. Combining clustering and co-training to enhance text classification using unlabelled data. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton: ACM, 2002. 620–625.

13 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

14 May C, Wang A, Bordia S, *et al*. On measuring social biases in sentence encoders. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 622–628.

15 Qiao Y, Xiong CY, Liu ZH, *et al*. Understanding the behaviors of BERT in ranking. arXiv: 1904.07531, 2019.

16 Zhang TY, Kishore V, Wu F, *et al*. BERTScore: Evaluating text generation with BERT. arXiv: 1904.09675, 2019.

17 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 3982–3992.

18 Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 815–823.

19 Nigam K, McCallum AK, Thrun S, *et al*. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2–3): 103–134.

20 McLachlan GJ, Basford KE. Mixture Models: Inference and Applications to Clustering. New York: Dekker, 1988.